Software Proposal Document for Classification of Alzheimer's by DNA Analysis project

Dr.Ashraf Abdelraouf , Ahmed Samir, Fairuz Soufy, Omar Ehab, Eng. Lobna Shaheen ,Sara Elbedeawi , Prof. Lamiaa Nabil , Dr. Omar El-Demrdash , Dr. Nora El-Smnoudy , Dr. Rawan El-Kady

March 16, 2020

Abstract

Artificial Intelligence and Genomics are two rapidly growing disciplines, combining them gives us so many opportunities for improving healthcare treatments and our understanding of the genome. So we used them to determine the 4 clear stages of Alzheimer which are: Normal patient(stage A), Alzheimer gene carrier(stage B), early stage patient(stage C), and late stage patient(stage D).They are classified according to their DNA. The main challenge for us is determining the stages B and C by the Artificial Intelligence part that is used and stages A and D are determined easily.

1 Introduction

1.1 Background

Alzheimer's disease is a permanent brain disorder that slowly destroys memory, thinking skills and finally the ability to carry out the simplest tasks[1]. 50 million people worldwide are living with dementia in 2018 [2]. Alzheimer can be detected by testing the DNA. The DNA is like two long ropes attached together with proteins to form structures named chromosomes. There are 46 chromosomes in the body (23 pairs)[3]. Each chromosome contains a huge number of segments, called genes. There are four genes related to Alzheimer: Amyloid precursor protein (APP), Presenilin-1 (PS-1), Presenilin-2 (PS-2), and Apolipoprotein E-e4 (APOE4)[4]. Each of these genes is made of a sequence of four characters : Adenine(A) , Thymine (T), Cytosine (C), Guanine (G). The four characters are the building block of the gene by forming a sequence of groups of three from the four characters. Each of these genes has a constant sequence. If the sequence of the above mentioned genes are altered then the patient will suffer from Alzheimer's. Stages A and D are easily to be differentiated between but the challenge is to distinguish between stages B and C which we solve by the artificial intelligence part we use to provide the best solution to slow down the evolution of Alzheimer.

1.2 Motivation

Knowing which stage the patient is in helps us reduce the progression of Alzheimer's [5] but it has been proven to be troublesome to find data-sets for a DNA sequence because it's very expensive to extract DNA and analyze it . The main CS challenges that we faced are:

1. Computational Power: Human DNA is stored in big files which needs a powerful machine in order to process them in the system.
2. The data sets aren't easily accessed or found.
3. The project requires high GPU to process the DNA quickly and compute the necessary operation in our neural network.

1.3 Problem Definition

The existing classifications of Alzheimer is either the patient is healthy (stage A) or the patient is in the last stage of Alzheimer(stage D) [6] which makes our main target is to show two new phases which are Alzheimer gene carrier(stage B) and early stage of Alzheimer(stage C). The gene carrier(stage B) is a healthy patient at the moment but has inherited the disease from his family and doesn't have any symptoms of the disease. The early stage(stage C) is a stage where the patient starts to notice some symptoms on manifesting.

2 Project Description

2.1 Objective

This system is developed to take the patient's DNA and push it to the system and the system automatically shows which stage the patient is in due to the training that was made with other data-sets which classify people into four stage. Without this system it was either the patient is healthy(stage A) or he is in the last phase of Alzheimer(stage D). It's hard to differentiate between stages B and C because the difference between them is very small and it needs a lot of training to gain more accurate classification. It's expensive to extract the DNA from the blood sample because it's a new technique that has a lot of process to pass by, materials to be used and few specialists[7]. In the future it can be developed to take shorter time or shorter process which may lead to make it more cheap in the future.

2.2 Scope

The system is developed to reveal which stage of Alzheimer the patient is in and how to help him to slow down the evolution of his disease depending on the stage the patient is in.

1. **Normal (Stage A) :** Healthy person with no family history or Alzheimer.
2. **Gene Carrier (Stage B) :** Healthy person carrier to a recessive gene for the disease from a family member.
3. **Early stage patient (Stage C) :** Patient with dominant gene carrier, being diagnosed as early stage "mild".
4. **Late stage patient (Stage D) :** Confirmed Alzheimer patient in the severe stage "late".

2.3 Project Overview

The system works as follows. The user will load the sample from the patient into Genius Prime application, which is used to open specific file type types(.fna or .fa) which contain the entirety of the human DNA which has the chromosomes and it makes us access each chromosome. We use chromosomes 1,14,19, and 21 and export them to new files. These files are loaded to the system in order to be reprocessed by initially cutting only the desired part of each file that contains the probable defected genes subsequently the files will be striped of new lines , numbers and spaces. Finally the content of the files will be grouped into groups of 3 and then those groups will be mapped into one hot vectors in order to be later processed by the neural network and be classified.

3 Similar System Information

1. **Predicting cancer type from tumour DNA signatures.**[8] The researchers tried to know the cancer type more accurately to give the best course of treatment to the patient. Their goal was to know the cancer type more accurately than before. They collected sequenced tumour DNA from Cancer Genomes. Around 6640 tumor samples showing 28 cancer types and used linear support vector machines with feature selection to predict the cancer type. They found that linear support vector machine is the most accurate model to predict cancer type with accuracy 49.4%. We saw how they used machine learning techniques to predict the cancer type.

2. **Convolutional neural networks for classification of alignments of non-coding RNA sequences** [9]. The researchers wanted to prove that Convolutional neural networks (CNN) is a good way for RNA analysis. The main problem statement is to classify non-coding RNA sequences into positive and negative classes to prove it's classifying correctly. The CNN classified the pairwise alignments of sequences for accurate clustering of sequences and show the benefits of the CNN method of inputting pairwise alignments for clustering of non-coding RNA (ncRNA) sequences and for motif discovery. The researchers solved a problem very similar to ours with a method analogous with the one that we intended to use demonstrating the feasibility of making the system. The accuracy of this project is 94.5%.

3. **Convolutional Neural Networks In Classifying Cancer Through DNA Methylation** [10]. The researchers decided to pursue the topic because traditional methods of cancer identification are generally not efficient. Moreover they usually require effort and have lower accuracy. The main problem is that the regular methods of cancer detection are quite troublesome moreover the possibility of false positives is present so a method with a higher accuracy was needed. The contribution the research team accomplished was building a model that can learn the changing DMR patterns to detect 32 cancer types. The model was able to attain a training accuracy of 96.54% and a testing accuracy of 92.87% the model was based on 10000 samples.

4. **Recurrent Convolutional Neural Networks for Text Classification** [11].Researchers wrote this paper to use Recurrent convolutional neural network (RCNN) for text classification. The key problem in text classification is feature representation. The reseachers used four separate text datasets to perform CNN and RCNN. Thus, they discovered that Neural Networks can collect more contextual feature data than conventional BoW-based approaches. This paper is important for my project as we may use the Recurrent Convolutional Neural Network (RCNN) for text classification.

5. **DNA Sequence Classification by Convolutional Neural Network** [12] The motivation of this paper is to prove that CCN has an excellent

performance in many fields even dealing with A, C, T and genes of the DNA. The main problem statement of the work is that DNA sequences are sequences of successive letters without space. There is no term of word in DNA sequence. so, we made a method to translate DNA sequences to sequence of words in order to apply the same representation technique for text data without losing position. The researches reached that the lowest improvement is nearly 1% of accuracy and the highest improvement is over 6% of accuracy. These improvements are quite high in comparison with other approaches such as finding good representations for sequences. This paper was important and valuable because it added to my knowledge in the sense that it showed how computer science and neural networks can help in diagnosing more than just one disease moreover it mentioned a plethora of guides and tips that may help in the proposed system.

3.1 Comparison with Proposed Project

1. This software is looking at genes APP,APOE,APSEN1,APSEN2 genes to see if they have a different sequence or are defected in comparison with normal genes to predict Alzheimer stage [13], while the similar system searches for tumor-ed DNA in the human cell.

2. The referenced system is using CNN to classify the RNA into two classes (positive class and negative class) while the proposed system is using it to classify some genes into 4 stages.

3. The proposed system was like a guide since it counters a similar problem. In comparison the systems are serving different purposes as the proposed system classifies and detects Alzheimer's into 4 classes but the referenced system classifies Different types of cancer (32). The proposed system classifies the patients based on changes in the DNA sequence but the referenced classifies based on DNA methylation.

4. The referenced project not only uses RCNN but also RNN and CNN therefore the project may shed some light on the pros and cons of each method of text classification. However referenced system focus is on NLP while the proposed system focus is classifying DNA sequences.

5. Their study about extracted features from the model's convolutional layers to see if there are any interesting and meaningful features extracted they are more concerned about the CNN as a classifier and they are not looking at specific genes but we are looking for genes that may be mutated so they may cause Alzheimer's diseases.

4 Project Management and Deliverables

4.1 Tasks and Time Plan

4.2 Budget and Resource Costs

1- The system needs Genius Prime Application because it is used in our project to open DNA sequence (.fna) files and convert it into the chromosome file type (.fasta) and (.gp) in order to be processed by the CNN we're using. It is 200$ per year for student license and 450$ for government and non-profit organizations to use.
2- The system needs GEFORCE GTX1080ti GPU to make the training part faster and more efficient because of its performance per dollar advantage over other Graphical processing units and the fact that it has a plethora of tensor cores over a high memory bandwidth which helps in many machine learning algorithms including CNN and RCNN. It is for 600$.

4.3 Supportive Documents

No surveys or reports have been carried out at the time of writing this text.

References

[1] H Förstl. What is alzheimer's disease. *Dementia*, 2:371–382, 2010.

[2] Christina Patterson. World alzheimer report 2018: the state of the art of dementia research: new frontiers. *Alzheimer's Disease International (ADI): London, UK*, 2018.

[3] I Cross and J Wolstenholme. An introduction to human chromosomes and their analysis. *Human Cytogenetics: constitutional analysis, 3rd ed. Oxford University Press*, 1:1–32, 2001.

[4] National Institute on Aging, NormanR Relkin, Alzheimer's Association Working Group, et al. Apolipoprotein e genotyping in alzheimer's disease. *The Lancet*, 347(9008):1091–1095, 1996.

[5] Daniel J Glass and Steven E Arnold. Some evolutionary perspectives on alzheimer's disease pathogenesis and pathology. *Alzheimer's & Dementia*, 8(4):343–351, 2012.

[6] Reisa A Sperling, Paul S Aisen, Laurel A Beckett, David A Bennett, Suzanne Craft, Anne M Fagan, Takeshi Iwatsubo, Clifford R Jack Jr, Jeffrey Kaye, Thomas J Montine, et al. Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):280–292, 2011.

[7] Diego Chacon-Cortes, Larisa M Haupt, Rod A Lea, and Lyn R Griffiths. Comparison of genomic dna extraction techniques from whole blood samples: a time, cost and quality evaluation study. *Molecular biology reports*, 39(5):5961–5966, 2012.

[8] Kee Pang Soh, Ewa Szczurek, Thomas Sakoparnig, and Niko Beerenwinkel. Predicting cancer type from tumour dna signatures. *Genome medicine*, 9(1):104, 2017.

[9] Genta Aoki and Yasubumi Sakakibara. Convolutional neural networks for classification of alignments of non-coding rna sequences. *Bioinformatics*, 34(13):i237–i244, 2018.

[10] Soham Chatterjee, Archana Iyer, Satya Avva, Abhai Kollara, and Malaikannan Sankarasubbu. Convolutional neural networks in classifying cancer through dna methylation. *arXiv preprint arXiv:1807.09617*, 2018.

[11] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[12] Ngoc Giang Nguyen, Vu Anh Tran, Duc Luu Ngo, Dau Phan, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Mamoru Kubo, and Kenji Satou. Dna sequence classification by convolutional neural network. *Journal of Biomedical Science and Engineering*, 9(05):280, 2016.

[13] David H Small and Catriona A McLean. Alzheimer's disease and the amyloid β protein: what is the role of amyloid? *Journal of neurochemistry*, 73(2):443–449, 1999.

Figure 1: System Overview

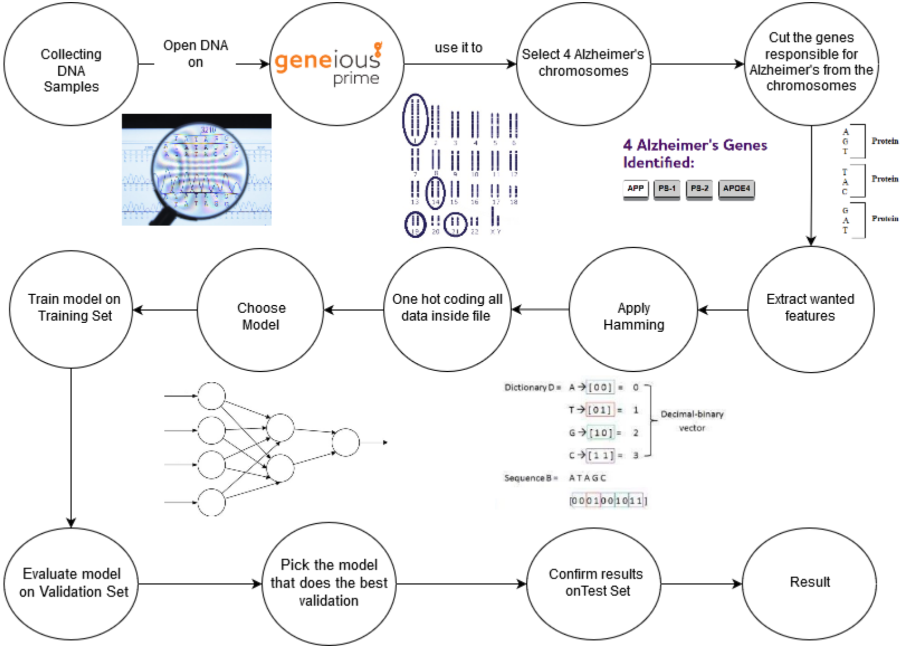
Collecting DNA Samples
Open DNA on
geneious prime
use it to
Select 4 Alzheimer's chromosomes
Cut the genes responsible for Alzheimer's from the chromosomes
4 Alzheimer's Genes Identified:
APP
PS-1
PS-2
APOE4
A G T
Protein
T A C
Protein
G A T
Protein
Extract wanted features
Apply Hamming
One hot coding all data inside file
Choose Model
Train model on Training Set
Dictionary D = A → [0 0] = 0
T → [0 1] = 1
G → [1 0] = 2
C → [1 1] = 3
Decimal-binary vector
Sequence B = A T A G C
[0 0 0 1 0 0 1 0 1 1]
Evaluate model on Validation Set
Pick the model that does the best validation
Confirm results onTest Set
Result

Figure 2: Timeline of work

Task	From	To
Collecting data-sets of stage A	30/9	5/10
Writing paper, presentation, and prototype	2/10	5/10
Collecting data-sets of stage D	15/10	20/10
Training the computer on differentiating between stage A and D	15/10	30/10
Collecting data-sets of stage B and C	30/10	15/11
Writing Survey Paper	15/11	1/12
Training the computer on differentiating between all stages	20/11	10/12
Writing SRS	1/12	7/12
Writing SDD	1/2	14/2
Finishing last paper	15/2	30/2
Validation and testing	30/2	15/3